強化学習におけるエージェントの行動選択に関する研究

強化学習におけるエージェントの行動選択に関する研究

STUDIES ON AGENT ’ S POLICY IN REINFORCEMENT LEARNING

高畑　慶
Kei TAKAHATA

指導教員 三浦孝夫

法政大学大学院理工学研究科システム理工学専攻修士課程

Reinforcement Learning (RL) is a learning method in which any agent learns from interaction with its environment. It allows us to acquire knowledge without any training data. However, for learning it takes time. In this work, we propose new method in agent's policy for efficient RL. In experiments, we show the effectivities of our approaches by some experimental results.

Key Words *: Reinforcement Learnin, Q-Learning, Kalman Filter, Retrospective Kalman Filter, Reverse Action Learning*

1.　問題の背景

自動運転や自律ロボットなどに強化学習が利用されている．強化学習では学習する主体をエージェント，エージェントと相互作用する対象を環境という．エージェントは状態の知覚と行動選択を行う．環境はエージェントの選択した行動に応じて正負を含む報酬を与え，エージェントを次の状態に遷移させる．この一連の流れをエージェントと環境の相互作用という．強化学習では，エージェントが環境と相互作用を行い知識を獲得する．エージェントが，環境から受け取る報酬の総和を最大にする知識を獲得することが強化学習の目的である．

強化学習は，教師あり学習と異なり明示的な教師データが存在しない．そのため，教師あり学習のように事前に大量の教師データを集める必要はない．また強化学習を利用することで，エージェントに事前ルールを付与する必要がなくなる．しかし，事前情報がないため，学習に時間がかかるという問題点がある．また，環境が定常であるとは限らないため環境の変化と共に知識を更新し続ける必要がある．

報酬の総和を最大にするために，エージェントは今までの経験から得た“知識の利用”と今より良い政策を見つけるための“探査”が必要となる．今までの経験により，報酬の総和が最大になると思われる行動のみを行う場合はより良い行動を見つけられない．つまり，“知識の利用”と“探査”は互いにトレードオフの関係にある．エージェントは両者をバランスよく行い，報酬の総和を最大にする知識の獲得をしなければならない．既存の行動選択手法も複数あるが，どのような行動選択手法が良いかは明らかになっていない．

本研究では強化学習の学習効率を改善するためのエージェントの行動選択手法を提案する．これにより，既存手法より少ない学習回数で知識の質を上げられる事を示す．

2.　扱う問題

(1)　カルマンフィルタを利用した Q 学習

強化学習は，学習データを用いず，環境からの報酬で知識を自動的に抽出できる．しかし，学習に時間がかかるという問題点がある．観測値から状態を予測するフィルタリング手法としてカルマンフィルタが存在する．カルマンフィルタは状態予測を行い，制御に利用されるのが一般的である．

本研究では，エージェントの行動選択にカルマンフィルタの状態予測を活用する手法を提案する．行動選択にカルマンフィルタの状態予測を利用することで，強化学習の学習効率を改善する．強化学習の代表的なタスクである追跡問題で実験を行い，提案手法の有効性を示す [1].

(2)　Q-Learning as Failure

学習効率改善のために，過去の行動の失敗を知識の改善に有効に活用する研究は少ない．本研究ではエージェントに失敗条件を定義し，失敗条件を満たした場合それまでに行った行動より逆の行動の方がベターだという仮定を置く．この仮定の下，逆行動学習を行う事で過去の失敗行動を知識の改善に有効活用することで学習の効率化を行う．QLKF を利用し，逆行動学習を行うためにカルマンフィルタの分散を戻す遡及的カルマンフィルタを定義し利用する．強化学習の代表的なタスクである追跡問題で実験を行い，提案手法の有効性を示す [2].

(3)　カルマンフィルタを利用したロバスト Q 学習

以前の研究では逆行動学習を行う場合，事前状態誤差の共分散行列を保持する必要があった．本研究では余分なメモリを使用せず，逆行動学習を行う手法を提案する．また，カルマンフィルタの観測誤差の分散が大きい場合でも逆行動学習により知識の改善が行える点を示す．本稿では協調距離問題というタスクを定義し，提案手法の有効性を示す [3].

表 1 収束回数

	ハンターのみ学習	ハンターと獲物が学習
ベースライン	16112	15589
提案手法	12888	12593

表 2 捕獲までのステップ数の比較（ハンタのみ学習）

学習時	QL	QL	QLKF	QLKF
評価時	QL	QLKF	QL	QLKF
捕獲までのステップ数の総和	78740	58851	35440	29725
比率	1	0.75	0.45	0.38

3. 結果

はじめにカルマンフィルタを Q 学習の実験を行う．評価の指標として 2 つの指標を使用する．1 つ目に，収束するまでの学習回数を表す収束までの学習回数である．回数が少ないほど速く学習できていると評価する．獲物が学習せず，ランダムに行動する場合と QL で学習する 2 パターンで実験を行う．2 つ目に，学習の経過ごとの捕獲までのステップ数である．これは少ないほど獲物を早く捕獲しているので知識の質が良いと評価する．ハンタのパターンとしては，学習時に QL(比較手法) か QLKF(提案手法) かの 2 パターンと，評価実験時に QL か QLKF かの 2 パターンで合計 4 パターンに分けられる．獲物は，学習してない場合はランダムな行動選択，QL で学習している場合は QL で行動選択を行うように設定し，捕獲ステップ数の評価実験を行う．(獲物は 2 パターンに分けられるので，全部で $4 \times 2 = 8$ パターンの実験を行う．) 収束回数の結果を表 1 に示す．獲物が学習しない場合の捕獲までのステップ数の結果を表 2，獲物が学習する場合の結果を表 3 に示す．

結果から，提案手法の収束回数は，Q 学習を行った時の約 0.8 倍に向上することがわかる．また捕獲までのステップ数は，学習時に QLKF を用いることにより，総ステップ数は通常の QL の半分以下のステップになることがわかる．

次に Q-Learning as Failure の実験を行う．追跡問題で実験を行い，捕獲までのステップ数を評価の指標に用いる．結果から QLRKF(提案手法) の捕獲までのステップ数の総数は獲物が学習しない場合，QL で学習する場合の 6 割 5 分，QLKF で学習する場合の 8 割 5 分に向上することががわかる．獲物が学習する場合，QL で学習する場合の 4 割，QLKF で学習する場合の 8 割 4 分に向上することがかわる．

最後に，カルマンフィルタを利用したロバスト Q 学習の実験を行う．2 体のエージェント (A,B) を用意し，協調距離問題で実験を行う．評価の指標として，協調距離内のステップ数を用いる．ステップ数が多いほど協調知識が獲得できていると評価する．結果から，学習 10 万回時点で QLRKF(提案手法) は QL の 342%，QLKF の 7% の改善率になることがわかる．また学習初期の学習 1 万回時点では QLKF の 286% の改善率になることがわかる．

4. 結論

本研究では強化学習の学習効率を改善するための行動選択手法を提案した．具体的に，強化学習の行動選択にカルマンフィルタの状態予測と失敗行動を利用する手法を提案し，効率的に学習が行わることを示した．

表 3 捕獲までのステップ数の比較（ハンタと獲物が学習）

学習時	QL	QL	QLKF	QLKF
評価時	QL	QLKF	QL	QLKF
捕獲までのステップ数の総和	176330	123461	36917	31109
比率	1	0.70	0.21	0.18

まず，カルマンフィルタを利用した Q 学習では，行動選択にカルマンフィルタの状態予測を利用する手法を提案した．提案手法の収束回数は，Q 学習を行った時の約 0.8 倍に向上することを示した．また捕獲までのステップ数の総数は，通常の QL の半分以下のステップになることを示した．

Q-Learning as Failure では，学習中の失敗行動を有効に活用する手法を提案した．QLKF を利用し，逆行動学習を行うためにカルマンフィルタの分散を戻す遡及的カルマンフィルタを定義し利用した．提案手法の捕獲までのステップ数の総数は獲物が学習しない場合，QL で学習する場合の 6 割 5 分，QLKF で学習する場合の 8 割 5 分に向上することを示した．獲物が学習する場合，QL で学習する場合の 4 割，QLKF で学習する場合の 8 割 4 分に向上することを示した．

カルマンフィルタを利用したロバスト Q 学習では，余分なメモリを使用せず，逆行動学習を行う手法を提案した．また，カルマンフィルタの観測誤差の分散が大きい場合でも逆行動学習により知識の改善が行える点を示した．学習 10 万回時点で QLRKF(提案手法) は QL の 342%，QLKF の 7% の改善率になることを示した．また学習初期の学習 1 万回時点では QLKF の 286% の改善率になることを示した．

参考文献

1) Kei Takahata, Takao Miura.; “Reinforcement Learning using Kalman Filters.” IEEE International Conference on Cognitive Informatics and Cognitive Computing(ICCICC). 2019
2) Kei Takahata, Takao Miura.; “Q-Learning as Failure.” THE 30TH INTERNATIONAL CONFERENCE ON INFORMATION MODELLING AND KNOWLEDGE BASES(EJC). 2020
3) ”カルマンフィルタを利用したロバスト Q 学習”, 第 13 回データ工学と情報マネジメントに関するフォーラム (DEIM), 2021